Variation in Universal Dependencies annotation : A token-based typological case study on adpossessive constructions

Sinnemäki, Kaius

The Association for Computational Linguistics
2020

Sinnemäki , K & Haakana , V L J 2020 , Variation in Universal Dependencies annotation : A token-based typological case study on adpossessive constructions . in M-C de Marneffe , M de Lhoneux , J Nivre & S Schuster (eds) , Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020) . The Association for Computational Linguistics , Stroudsburg , pp. 158-167 , Workshop on Universal Dependencies , 13/12/2020 . < https://universaldependencies.org/udw20/papers/2020.udw2020-1.18.pdf >

http://hdl.handle.net/10138/323330

Variation in Universal Dependencies annotation: A token-based typological case study on adpossessive constructions

Kaius Sinnemäki
General linguistics
P.O. Box 24 (Unioninkatu 40)
00014 University of Helsinki
FINLAND
kaius.sinnemaki@helsinki.fi

Viljami Haakana
General linguistics
P.O. Box 24 (Unioninkatu 40)
00014 University of Helsinki
FINLAND
viljami.haakana@helsinki.fi

Abstract

In this paper we present a method for identifying and analyzing adnominal possessive constructions in 66 Universal Dependencies treebanks. We classify adpossessive constructions in terms of their morphological type (locus of marking) and present a workflow for detecting and analyzing them typologically. Based on a preliminary evaluation, the algorithm works fairly reliably in adpossessive constructions that are morphologically marked. However, it performs rather poorly in adpossessive constructions that are not marked morphologically, so-called zero-marked constructions, because of difficulties in identifying these constructions with the current annotation. We also discuss different types of variation in annotation in different treebanks for the same language and for treebanks of closely related languages. The research focuses on one well-circumscribed and universal construction in the hope of generating more interest in using UD for cross-linguistic comparison and for contributing towards developing yet more consistent annotation of constructions in the UD annotation scheme.

1 Introduction

Universal Dependencies (UD) and other multilingual language corpora have in many ways boosted cross-linguistic corpus research to a completely new level, enabling novel developments, for instance, in the so-called token-based typology (Levshina, 2019). One of the aims of UD is to provide universal annotations for various syntactic relations (called universal dependency relations in UD), parts-of-speech, and grammatical categories (features in UD), which then enable productive cross-linguistic comparison. However, the currently available universal dependency relations are what from a typological perspective could be described as higher-level functions, such as subject, object, nominal modifiers, and adjectival modifiers. In addition, the dependency relations are not consistently based on universal construction types, as observed and recommended by Croft et al. (2017), but often on language-specific strategies. For instance, nominal modifiers encompass constructions that have a wide range of functions in the world's languages, including possession, but it is not easy to compare them in the UD treebanks owing to variation in annotation across treebanks. Such variation has been noted earlier as well, in relation to research on linguistic complexity (Berdicevskis et al., 2018).

In this paper we focus on one well-defined and universal construction, namely adnominal possessives (also called adpossessive constructions or possessive noun phrases). Adpossessive constructions are syntactically noun phrases whose head is a noun and that may have a noun or a pronoun as a dependent modifier. Semantically the relation between the head and the dependent in these constructions expresses typically ownership (alienable possession), such as *my car*, part-whole relationships (including inalienable possession), such as *my hand*, and kinship relationships, such as *my daughter*. Syntactic adpossessive constructions can also be used for various other functions depending on each language (Koptjevskaja-Tamm, 2003; Haspelmath, 2017; Ortmann, 2018). In adpossessive constructions the syntactic **dependent** is semantically the possessor and the syntactic **head** is semantically the possessee. For

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 158–167
Barcelona, Spain (Online), December 13, 2020

example, in the construction *my daughter* the syntactic head is the possessee *daughter* and the dependent is the possessor *my*.

We aim to research how these constructions can be identified in the UD treebanks. Our analyses are based on a sample of 63 treebanks from 44 languages that have a large corpus in the UD. The selected languages represent ten language families from Eurasia, Africa, and the Pacific (see Table 1). The treebanks were part of a shared task in the Interactive Workshop on Measuring Language Complexity (IWMLC), organized in September 2019 at the Freiburg Institute for Advanced Studies (FRIAS) in Freiburg, Germany. The treebanks were pre-selected by the workshop organizers and the analyses were based on Universal Dependencies (UD) 2.3 corpora (Nivre et al., 2018). Our work for identifying adpossessive constructions started in relation to that workshop, but the full results are published here for the first time. Our analyses, algorithms, and the full dataset are published as supplements in the hope of encouraging and enabling further development of universal annotations in the UD treebanks.[1]

Family	**N**	**%**
Afro-Asiatic	2	4.5
Austro-Asiatic	1	2.3
Austronesian	1	2.3
Basque	1	2.3
Indo-European	31	70.5
Japanese	1	2.3
Korean	1	2.3
Sino-Tibetan	1	2.3
Turkic	2	4.5
Uralic	3	6.8

Table 1: Distribution of sample languages.

In the following section, we discuss the process of identifying adpossessive constructions from the UD treebanks. In Section 3 we discuss several examples of variation in the UD annotation, followed by a short discussion and conclusions in Section 4.

2 Identifying and classifying adpossessive constructions

2.1 Preliminaries

In token-based typological research, constructions can ideally be identified using a small set of criteria across treebanks. For instance, the dependency relation `nsubj` identifies nominal subjects consistently in all UD treebanks. Further division into subjects of intransitive and transitive predicates is less straightforward to do but possible by identifying those clauses in which the predicate has a dependent with the dependency relation `obj`.

As for adpossessive constructions, the treebanks may optionally use the dependency relation `nmod:poss` (or `det:poss`). These relations are subtypes, which are not universally defined and their usage thus varies depending on language-specific criteria. In many treebanks, `nmod:poss` is used for distinguishing non-adpositional possessives from adpositional possessives. For instance, in the DDT treebank for Danish (see the example in Table 2), the subtype `nmod:poss` is used for non-adpositional possessives. In adpositional possessives, such as in the Danish *af* construction (Table 3), the dependency relation of the possessor is coded as `nmod`. Danish thus uses two types of adpossessive constructions and in the latter type it is necessary to identify the construction using information about the adposition, because the subtype `nmod:poss` is not used. Such practice results in a wider range of annotations compared to, for instance, the identification of the nominal subject and object. Note that in some treebanks, such as the CRB treebank for Bambara (not in our sample), the relation `nmod:poss` is used for

[1]The supplements are available at <https://version.helsinki.fi/gramadapt/udw2020-adpossessive-constructions>. Note that since our purpose is to identify adpossessive constructions, the algorithms are aimed for that purpose and not, for instance, for fixing UD annotations.

both adpositional and non-adpositional possessives. This would be recommendable for other treebanks as well, because adpositional possessives can be distinguished from non-adpositional possessives in any case by using the possessive adposition as a separate criterion.

N	Wordform	Lemma	UPOS	Features	Head	Dependency
23	i	i	ADP	AdpType=Prep	25	case
24	Camillas	Camilla	PROPN	Case=Gen	25	nmod:poss
25	sofa	sofa	NOUN	Definite=Ind|Gender=Com|Number=Sing	22	nmod

Table 2: An adpossessive construction in the Danish treebank DDT (sent_id = dev-268; excluding some columns).

N	Wordform	Lemma	UPOS	Features	Head	Dependency
16	i	i	ADP	AdpType=Prep	17	case
17	form	form	NOUN	Definite=Ind|Gender=Com|Number=Sing	7	obl
18	af	af	ADP	AdpType=Prep	19	case
19	udgifter	udgift	NOUN	Definite=Ind|Gender=Com|Number=Plur	17	nmod

Table 3: An adpositional adpossessive construction in the Danish treebank DDT (sent_id = dev-100; excluding some columns).

2.2 Locus of marking

As implied by the discussion of the Danish *af* construction, it may not be possible to identify adpossessive constructions without information about how the dependency relation between the head and the dependent in these constructions is marked morphologically. In cross-linguistic research the morphological typology of this relation is called locus of marking (Nichols, 1992). Locus of marking refers to the position of morphological marking of syntactic relations (or dependencies) in a construction. There are four logical loci for morphological marking, illustrated here for adpossessive constructions: it occurs either on the head of the construction (the possessed in possessive NP) as in (1a), the dependent of the construction (the possessor) as in (1b), on both (called double marking) as in (1c), or on neither (called zero marking) as in (1d).[2]

(1) a. Head marking (Indonesian, Austronesian; (Sneddon, 1996, 146))

ibu-nya Suparjo
mother-3SG.POSS Suparjo

'Suparjo's mother'

b. Dependent marking (Swedish, Germanic, Indo-European)

min bok
1SG.POSS book

'My book' (Swedish)

c. Double marking (Finnish, Finnic, Uralic)

häne-n pyörä-nsä
3SG-GEN bike-3POSS

'his bike'

d. Zero-marking (Indonesian, Austronesian; (Sneddon, 1996, 144))

rumah Tomo
house Tomo

'Tomo's house'

[2]We exclude a fifth type called floating marking, which is cross-linguistically rare.

Locus of marking may vary across constructions even within the same language and certainly across languages. This means that one language may have several different types of adpossessive constructions depending on the particular language-specific strategies of marking the dependency relation (e.g., genitive case, adpositions, and possessive suffixes) as well as on their locus of marking. In the absence of systematic and universal annotation for adpossessive constructions, their identification process is essentially driven by identifying treebank-specific morphological strategies and their loci from the corpora (Croft et al., 2017). That process is very similar to the regular work that typologists practice in cross-linguistic comparison.

2.3 Workflow

Typological research starts by defining the object of research, called comparative concepts by Haspelmath (2010), in such a way that similar constructions can be identified and compared in the sample languages. Constructions and patterns in particular languages are then analyzed in relation to the comparative concept and classified to different types. This is what we did for identifying adpossessive constructions in the UD treebanks. We define adpossessive constructions by using both morphosyntactic and semantic criteria. Semantically adpossessive constructions express a possessive relationship, such as ownership, kinship relation, or part-whole relationship; morphosyntactically they are noun phrases (sometimes single nouns) whose syntactic head noun functions as a possessee and that may also be modified by a noun or a bound form that functions as a possessor (Koptjevskaja-Tamm, 2003; Haspelmath, 2017).[3] When identifying adpossessive constructions in the UD treebanks, our workflow was roughly as described schematically in Figure1.

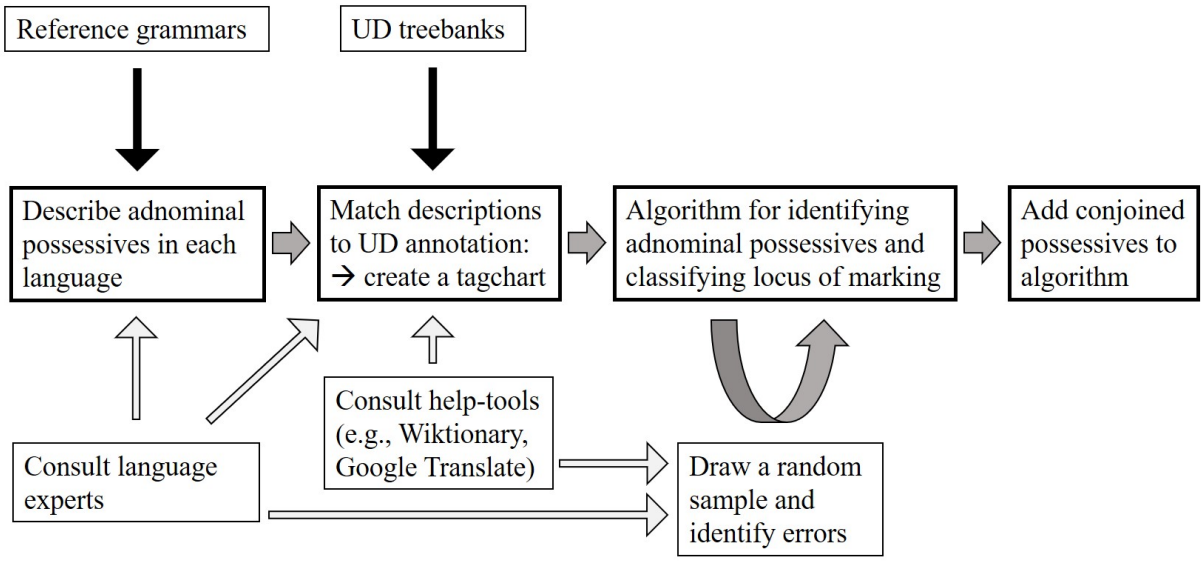

Figure 1: Workflow for identifying adpossessive constructions.

First, we used reference grammars, such as Schmidt (1999) on Urdu, to describe how adpossessive constructions are marked in the sample languages. The reference sources that we used are listed at the end of Supplement S1. Next we matched these descriptions to the annotation of adnominal possessives in the UD treebanks. For the majority of treebanks (45 out of 63) it was possible to create a tagchart of the treebank-specific annotations for identifying adnominal possessives in their various forms. Table 3 lists the 22 UD tags used in the tagchart. In addition to these tags, dozens of individual constructions were identified by comparing the lemmas and surface forms to one another. For the first two steps in the workflow we consulted language experts as well as online tools, such as Wiktionary and Google Translate, to help us determine how adpossessive constructions are annotated in the UD. Third, we wrote Python

[3]Note that we use a somewhat broader definition of possessive relationship than e.g. Koptjevskaja-Tamm (2003) and Haspelmath (2017), who delimit possessive relationship only to ownership, kinship relation, and part-whole relationship.

algorithms to detect each individual adpossessive construction in the 63 treebanks. For 18 treebanks we created separate algorithms because they were impossible to analyze by merely listing a set of required annotations in a tagchart. Supplement S2 contains the tagchart, the Python algorithms, and some other files involved in the computational analysis. Fourth, we randomly selected a few dozen sentences containing adnominal possessives from each treebank, detected errors in the analysis, doublechekced that the identified constructions expressed a possessive relationship, and updated the algorithm accordingly. Emphasis was on updating and correcting the algorithm rather than producing evaluative data on its performance. Fifth, we classified locus of marking for each individual adpossessive construction. This step was done largely in parallel with step three. As the last step, the algorithm was updated to include conjoined possessors where possible. The output of this process is a dataset (Supplement S3) that contains information about each individual adpossessive construction that we identified and classified. The analyzed UD treebanks (the `CoNLL-U` files) are available at the website of the IWMLC workshop.[4] Our analyses follow the principles of late aggregation in the spirit of Levshina (2019) and Zakharko et al. (2017), which enables researching language-internal variation at the level of each individual dependency relation without needlessly aggregating the data in different ways.

We delimited the analysis of adpossessive constructions in a few important ways. First, we focus on constructions in which the possessed is a full noun, that is, either a common noun (POS tag `NOUN` or a proper noun (POS-tags `PROPN`). Second, we focus on constructions in which the possessor is a personal possessive pronoun (e.g., feature `PronType=Prs`), a common noun, a proper noun, or a possessive adjective (as in many Slavic languages). However, we generally exclude demonstrative and other pronouns as possessors. Third, we define head/dependent marking as any morphological marking on the head or the dependent, be it via affixes, tones, morphophonological alternations, clitics, or independent function words (Bickel and Nichols, 2007). In the UD annotation the morphological marking is identifiable from the feature annotation and from the use of separate function words, such as particles and adpositions, for modifying the possessor. Relevant annotations include features such as `Case=Gen` and `Poss=Yes` for identifying dependent marking and layered features, such as `psor` for identifying head marking (see Table 4). In addition, we used information about particles and adpositions that mark the possessor as the dependent of the adpossessive construction (see Table 1).

Type of annotation	Tags
Parts-of-speech tags	ADJ, ADP, DET, NOUN, PART, PRON, PROPN
Feature annotations	Case, Number, Person, Poss, PronType, psor, Reflex
Dependency relations	amod, case, case:gen, det, det:poss, nmod, nmod:att, nmod:poss, nmod:gobj, nmod:gsubj

Table 4: UD annotations used in the identification and classification of adpossessive constructions.

2.4 Zero marked adpossessives

Adpossessive constructions with zero marking were probably the most difficult ones to identify and we expect most of the unresolved issues to concern this type. The most important reason for this was that in many languages it was necessary to use annotation about morphological marking to identify adpossessive constructions to begin with. As a result, zero-marked constructions were sometimes practically impossible to detect reliably. For instance, in adpossessive constructions in Vietnamese the dependency relation is marked with a possessive preposition (Thompson, 1965). However, this preposition is optional in some contexts, leading to zero marked adpossessive constructions. Yet it was difficult to identify adpossessive constructions reliably without recourse to the possessive preposition, because that would have meant classifying all noun-noun juxtapositions as adnominal possessives, which seemed to result in many false positives. For this reason, in the Vietnamese treebank we could sift only dependent marked adpossessive constructions (with noun possessors) but not zero marked ones.

For research in token-based typology zero-marked adnominal possessives are, however, theoretically very interesting at least for two reasons. For one, in many languages the dependency relation in in-

[4]The selected UD treebanks can be accessed at <http://www.christianbentz.de/MLC2019/UDtrack.zip>.

alienable possessives is morphologically zero marked but in alienable possessives it tends to be overtly marked. While there is ongoing debate about the reasons for this typological distribution, frequency and predictability may be among the strongest factors causing it (Haspelmath, 2017). In addition, from the perspective of linguistic efficiency it would be natural to hypothesize that zero marking would be preferred only when the head and the dependent are adjacent to one another and that the probability of morphological marking would increase as a function of dependency length (Gibson et al., 2019). These issues are exactly what multilingual annotated corpora, such as UD, are excellent tools for, but only to the extent they provide the sufficient means for reliably identifying the relevant constructions across treebanks. This is something that seems currently problematic especially for zero marked adpossessive constructions.[5]

3 Results

Our algorithm identified altogether 724 694 adpossessive constructions in the data. The distribution of different morphological types across sample languages are presented in Figure 2 as percent shares. Dependent marking is clearly a dominating pattern overall. It occurs in all but four languages, occurs with at least 65% share in 39 languages, and is the only type in 18 languages. This strong domination of dependent marking is an areal feature. Head and dependent marking in adpossessive constructions are fairly evenly distributed in the world's languages, but dependent marking dominates in Eurasia and Africa and head marking in the Americas (Nichols and Bickel, 2013). In our data head marking is a relatively minor type, occurring only in six languages, but where it occurs it is a dominating pattern (in five languages with 65% share or more). These figures for head and dependent marking clearly reflect the fact that languages of Eurasia are overrepresented in the sample. Double marking is a rare and minor type occurring in only three languages (Finnish, Turkish, and Uyghur) and only as a minor type in each of them (with less than 20% shares in each). Zero marking, on the other hand, is a common pattern, occurring in 23 sample languages (52% of languages); however, it is only a minor pattern in most of these languages, being fairly common only in Indonesian and Estonian. In addition, zero marking seems to occur in languages that have also dependent marking, but in this sample this is clearly a side-effect of dependent marking dominating in adpossessive constructions overall.

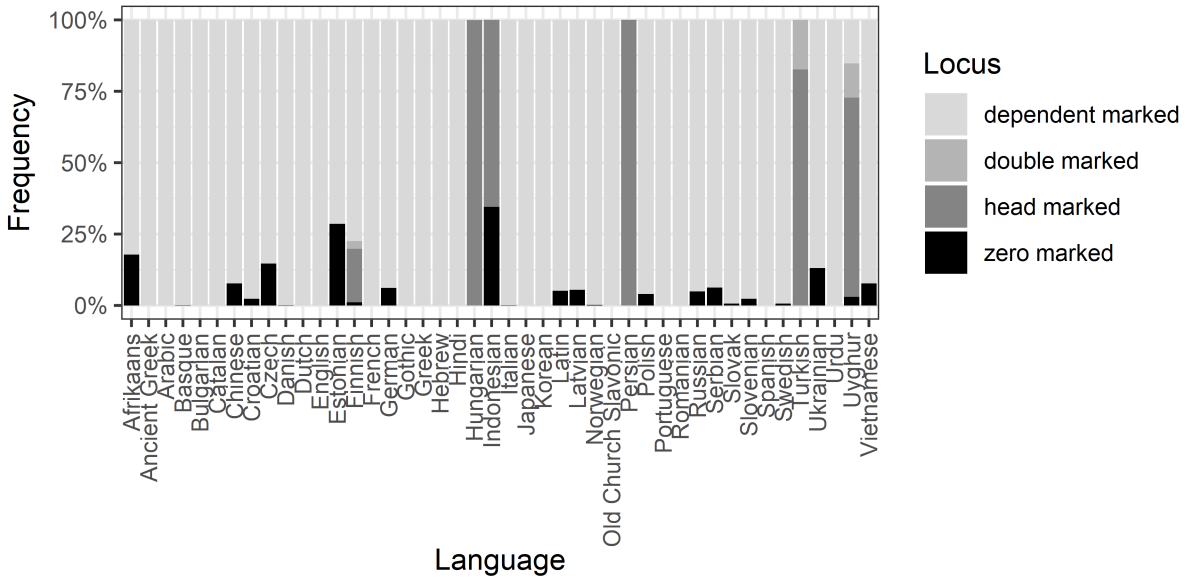

Figure 2: Distribution of morphological types in adpossessive constructions.

[5]Our dataset contains information also about length of dependency between the head and the dependent as well as their word order, as these were straightforward to compute and will increase the usefulness of the dataset in future research. Within the limited scope of this paper it is not possible to report on any results concerning these additional features.

Overall the process for identifying adpossessive constructions in the sampled treebanks required us to define up to 1,000 tags (on average 14 per treebank), including POS tags, specific features, dependency relations, and lemmas for pronouns and adpositions in several cases. Figure 3 presents a boxplot of how many tags were needed to identify adpossessive constructions in the treebanks, grouped into language families.[6] This figure represents only those languages included in the tagchart. The boxplot suggests that languages in most families required roughly 10-15 tags, in a few languages around 20 tags, and in the Turkic family more than 30 tags. This variation may reflect real diversity in adpossessive constructions in these language families, but it may at least in part reflect also the variation in the annotation of adpossessive constructions in the treebanks, suggesting potential places of concern in terms of annotation consistency.

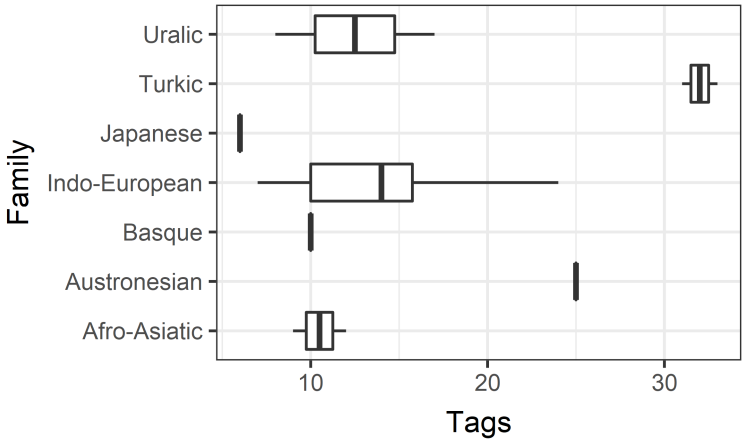

Figure 3: Number of tags needed in each language family for identifying adpossessive constructions.

It was not possible to estimate automatically the performance of the algorithm, including the range and incidence of false positives and negatives. The main reason for this is that there is no gold standard available, which precludes comparing our results against such benchmark. Whether such a benchmark will be reached in the future is an open question. However, to achieve at least a very preliminary and crude idea of the algorithm's performance, we manually analyzed a few dozen adpossessive constructions in four languages from four different language families. The results of this evaluation are presented in Table 5. Recall and precision of the algorithm were quite high in Afrikaans, Finnish, and Turkish (both ¿ 0.9). In these languages there were no false positives and only a handful of false negatives. In Indonesian, precision was 0.92, but recall was only 0.34, mostly due to a high number of false negatives. The algorithm for Indonesian currently fails to detect zero marked adpossessive constructions adequately: all false positives and negatives are of this type. This is no wonder, since a range of different functions (apparently at least `det`, `obj`, and `compound`) seem to be used for zero-marked possessors in the Indonesian GSD treebank, and it is not very clear to us whether any of these functions are systematically used for this construction. Overall, it seems that morphologically overtly marked adpossessive constructions can already be fairly reliably identified despite language-specific variation. More effort is needed to identify zero-marked adpossessive constructions in an equally reliable way.

Language	Recall	Precision	False positives	False negatives	Adpossessive constructions	Sentences
Afrikaans	0.96	0.92	0	2	52	30
Finnish	0.97	1.00	0	1	32	40
Indonesian	0.34	0.92	3	23	35	53
Turkish	0.94	1.00	0	2	33	30

Table 5: Evaluation of algorithm performance.

[6]The package `ggplot2` (Wickham, 2016) was used for graphics and statistics in the R environment (R Core Team, 2020).

The detailed linguistic analysis involved in identifying adpossessive constructions from the UD treebanks led to noting some unexpected variation in annotations, some of which we describe here. One issue concerns the differential treatment of clitics. For instance, in the English treebanks, the possessive clitic *'s* is analyzed as a separate token on its own. In Korean, the possessor is marked with the possessive clitic *-uy*, but in the Korean treebanks this clitic is analyzed neither as a separate token nor is its morphological category annotated in the features. We identified this clitic as the last element of the lemma, and analyzed all adpossessive constructions as dependent marked. In the Korean GSD treebank this clitic is marked in the language-specific POS as `JKG` and in the Korean Kaist treebank in the language-specific POS as `jcm`, so its identification could have been done in different ways. Another issue concerned Persian, in which the possessive relation is marked on the head via the so-called ezafe construction. However, ezafe is usually not at all marked in written language, and therefore it is not visible in the Persian Seraji treebank at all. We thus assumed that all identified adpossessive constructions had also the ezafe and they were accordingly classified as head marked.

Sometimes different treebanks for the same language varied a lot in their logic for annotation. As an example, in the ITTB treebank for Latin the POS of possessive pronouns is `ADJ`, its features contain `Poss=Yes`, and its dependency relation is `amod` or `nmod`. In the PROIEL treebank for Latin, on the other hand, the POS of possessive pronouns is `PRON`, its identifying features contain `Case=Gen` and `PronType=Prs`, and its dependency relation is `det`. Possessive pronouns in the Latin treebanks are thus identified with different tags on all three major criteria that we used (POS, features, dependencies).

There was variation also between treebanks for closely related languages. As an example, Slavic languages have two distinct adpossessive constructions, the possessive genitive and the possessive adjective constructions. In the majority of the sampled Slavic treebanks, the POS of the possessor is `ADJ`, its features contain `Poss=Yes`, and its dependency relation is `amod`. However, in the SNK treebank for Slovak, possessive adjectives cannot be identified and distinguished from other adjectives, because they are not tagged with the feature `Poss=Yes`. In other words, all possessive adjectives were unnoticed by our algorithm for Slovak.

Syncretism is another cause of variation in the treebanks. Because it was important for us to try detecting zero marking, for instance, via detecting absence of case distinctions, syncretism caused some issues with analyzing locus of marking. Table 6 presents an illustrative example from the German GSD treebank. Like many German nouns, the word *Region* does not inflect for case at all. However, according to the feature annotation *Region* is in the genitive case and thus using the feature `Case=Gen` as an identifying tag would have resulted in analyzing *Region* as dependent marked. On the other hand, because the article inflects for case, analyzing this construction as dependent marked would have been correct in any case. The example illustrates the fact that the annotated features do not necessarily reflect the surface structures but more abstract structures. One solution to syncretism would be to synthesize UD with Unimorph, which does address syncretism (McCarthy et al., 2017).

N	wordform	lemma	UPOS	features	head	dependency
8	der	der	DET	Case=Gen|Definite=Def|Gender=Fem|Number=Sing|PronType=Art	9	det
9	Region	Region	NOUN	Case=Gen|Gender=Fem|Number=Sing	7	nmod

Table 6: Example of an adpossessive construction in the German GSD treebank (sent_id = dev-s15); excluding unnecessary columns.

When determining whether the construction was zero marked we often compared the lemma of the possessor's form directly with its surface form. If the two forms were identical, we analyzed the adpossessive construction zero marked, otherwise as dependent (or double marked). However, because the difference between the lemma and the surface form may depend on many other features besides e.g. case marking, we limited the lemma comparisons to contexts in which the other features were identical. In other words we excluded comparing genitive plural and nominative singular when they were identical and different from nominative plural, for instance, and compared the case-inflected forms only in the singular.

The UD annotations sometimes have tags that do not reflect the word form in itself. These examples are probably very rare, but still worth mentioning. Consider the first two words of the sentence `b204.33` in the Finnish TDT treebank: *Meidän suhde* 'our relationship'. In Standard Finnish the head of this adpossessive construction has a possessive suffix, resulting in the word form *suhteemme* instead of the base form *suhde* which occurs in colloquial usage. However, the features for this word contained `Number[psor]=Plur` and `Person[psor]=1`, which represent the standard form with the possessive suffix. Such examples from colloquial usage are probably so difficult and inefficient to detect systematically that their existence in the treebanks have to be accepted.

4 Discussion and conclusion

Our analyses suggest that there are different types of variations in how adpossessive constructions are annotated across UD treebanks. To some extent this is expected because languages sometimes have several different morphosyntactic ways for expressing adnominal possession and across languages this diversity is multiplied. On the other hand, adnominal possession is a prominent and universal syntactic construction in the languages of the world, and reducing unnecessary variation in annotation would enable more efficient identification of these constructions by linguists and language technologists.

Given the variation in annotation it is natural that our method for identifying adpossessive constructions contained both false positives and false negatives, as even the limited evaluation suggested. False negatives were rare and they seem structurally quite similar with adpossessive constructions albeit expressing some other function which is difficult to delineate from adpossessive constructions based on the current annotation. False negatives, on the other hand, result largely from errors in the algorithm, from our insufficient knowledge of the languages and treebanks, and from problems with the current annotation of the treebanks, as discussed in relation to zero marking in Indonesian and Vietnamese. A further challenge, for instance, in Vietnamese, is possessive classifiers (Hui, 2005), but since we did not have sufficient knowledge about this type of construction, we did not even try detecting them. Overall, since in many languages morphological annotation had to be used for identifying adpossessive constructions, it is possible that in quite many languages potential zero-marked adpossessive constructions went unnoticed.

The success and usefulness of UD and other multilingual language corpora rest largely on their annotation schemes. In this paper we have reported on a typological case study of identifying and analyzing adpossessive constructions in 63 UD treebanks. This construction represent much variation in annotation even in different treebanks of the same language and in treebanks of closely related languages. In line with earlier research (Berdicevskis et al., 2018), our results suggest that there are some limitations to the extent which UD can currently be used for cross-linguistic research, especially concerning zero-marked constructions; however, the results also indicate that for morphologically marked adpossessive constructions, UD can currently be used quite reliably for token-based typological research.

Acknowledgements

We are grateful to the following people for help with linguistic analyses: Çağrı Çöltekin (Turkish), Sonja Dahlgren (Ancient Greek), Andrei Dumitrescu (Romanian), Yoonmi Oh (Korean), Marja Vierros (Ancient Greek), and Max Wahlström (Old Church Slavonic). All remaining errors are our own. We thank Miikka Silfverberg for help with UD-related technical issues and Jarmo Niemelä for help with technical issues related to ortographies using LaTeX. Some earlier parts of this research were presented at the Interactive Workshop Measuring Language Complexity (IWMLC), in Freiburg in September 2019 and at the Helsinki Area and Language Studies seminar in Helsinki in February 2020; we are grateful for the organizers of these workshops for having invited us, and for the participants for useful comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805371).

References

Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama, and Christian Bentz. 2018. Using Universal Dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17, Brussels, Belgium.

Balthasar Bickel and Johanna Nichols. 2007. Inflectional morphology. In Timothy Shopen (ed.) *Language Typology and Syntactic Description 3*, pages 169–240, Cambridge University Press, Cambridge.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic Typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75, Bloomington, IN, USA.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Martin Haspelmath. 2017. Explaining alienability contrasts in adpossessive constructions: Predictability vs. iconicity. *Zeitschrift für Sprachwissenschaft*, 36(2):193–231.

Sim Sook Hui. 2005. The Semantics and Grammar of Vietnamese Classifiers. MA Thesis, National University of Singapore.

Maria Koptjevskaja-Tamm. 2003. Possessive noun phrases in the languages of Europe. In Frans Planck (ed.) *Noun phrase structure in the languages of Europe*, pages 621–722, Mouton de Gruyter, Berlin.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium.

Johanna Nichols. 1992. *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago, IL.

Johanna Nichols and Balthasar Bickel. 2013. Locus of marking in possessive noun phrases. In Matthew S. Dryer and Martin Haspelmath (eds.) *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available online at http://wals.info/chapter/24.

Joakim Nivre, Mitchell Abrams, Željko Agić, et al. 2018. Universal Dependencies 2.3. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2895.

Albert Ortmann. 2018. Connecting the typology and semantics of nominal possession: alienability splits and the morphology–semantics interface. *Morphology*, 28(1):99–144.

R Core Team. 2020. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. Available at https://www.R-project.org.

Ruth Laila Schmidt. 1999. *Urdu: An Essential Grammar*. Routledge, London.

James N. Sneddon. 1996. *Indonesian: A Comprehensive Grammar*. Routledge, London.

Laurence C. Thompson. 1965. *A Vietnamese Grammar*. University of Washington Press, Seattle, WA.

Taras Zakharko, Alena Witzlack-Makerevich, Johanna Nichols, and Balthasar Bickel. 2017. Late aggregation as a design principle for typological databases. ALT Workshop on Design Principles of Typological Databases, 15 December 2017.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, NY.